Revisiting Feature Fusion for RGB-T Salient Object Detection

Zhang, Qiang; Xiao, Tonglin; Huang, Nianchang; Zhang, Dingwen; Han, Jungong

Published in:
IEEE Transactions on Circuits and Systems for Video Technology

DOI:
10.1109/TCSVT.2020.3014663

Publication date:
2021

Citation for published version (APA):
Zhang, Q., Xiao, T., Huang, N., Zhang, D., & Han, J. (2021). Revisiting Feature Fusion for RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(5), 1804-1818. [9161021]. https://doi.org/10.1109/TCSVT.2020.3014663

Revisiting Feature Fusion for RGB-T Salient Object Detection

Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang* and Jungong Han*

***Abstract*—While many RGB-based saliency detection algorithms have recently shown the capability of segmenting salient objects from an image, they still suffer from unsatisfactory performance when dealing with complex scenarios, insufficient illumination or occluded appearances. To overcome this problem, this paper studies RGB-T saliency detection, where we take advantage of thermal modality's robustness against illumination and occlusion. To achieve this goal, we revisit feature fusion for mining intrinsic RGB-T saliency patterns and propose a novel deep feature fusion network, which consists of the multi-scale, multi-modality, and multi-level feature fusion modules. Specifically, the multi-scale feature fusion module captures rich contexture features from each modality feature, while the multi-modality and multi-level feature fusion modules integrate complementary features from different modality features and different level of features, respectively. To demonstrate the effectiveness of the proposed approach, we conduct comprehensive experiments on the RGB-T saliency detection benchmark. The experimental results demonstrate that our approach outperforms other state-of-the-art methods and the conventional feature fusion modules by a large margin.**

***Index Terms*—Salient object detection, RGB-T, Multi-scale, Multi-modality, Multi-level, Feature fusion**

I. Introduction

SALIENT object detection aims to detect objects that can attract human attention in various scenarios and accurately segment object and background regions. Due to its high research value and significance in practical applications, salient object detection has attracted a lot of attention from the fields of image processing, computer vision, pattern recognition and artificial intelligence. Recent studies have shown that the salient object detection techniques can be applied in many applications, such as object recognition [1], image and video compression [2], image segmentation [3], content-based image resizing [4], visual tracking [5], image retrieval [6] and person re-identification [7].

RGB-based saliency detection models have been well studied for a long time and achieved great progress [8]–[16]. Conventional RGB saliency models can be roughly divided into two categories [17]: top-down [16], [18], [19] and bottom-up [15], [20], [21] pipelines. Top-down models are primarily based on high-level saliency priors of a particular category and various hand-crafted features, so that they are usually task-driven models. Bottom-up models, however, require low-level visual features such as color, texture, and contrast. Generally, these traditional methods depend mainly on hand-crafted saliency features and may be unreliable in discriminating the salient objects from complicated scenarios due to the lack of high-level contexts. In recent years, the Fully Convolutional Neural Network (FCN) has been widely used in many computer vision tasks [22], [23] and a large number of FCN-based models have been designed for RGB-based salient object detection. By building multi-level and multi-scale feature representations, these FCN-based salient object detection models achieved appealing performance.

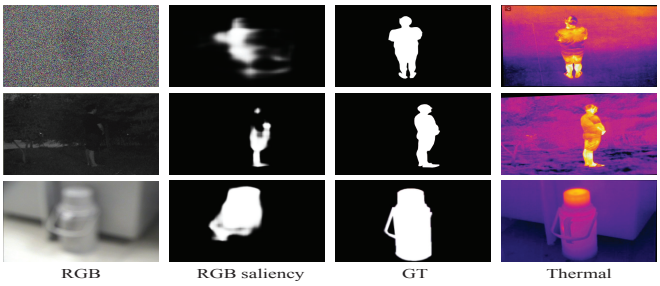


Fig. 1. Saliency maps generated by RGB-based salient object detection model for some images in complex scenarios (occlusion, lack of illumination and low contrast) and their ground truth annotations. The last column shows the corresponding spatially aligned thermal images to the RGB images.

Most of those saliency detection models perform saliency detection merely on RGB images, which making use of detailed visual cues, such as color, texture, and spatial details. However, such visual cues are sensitive to illumination, weather condition, and occlusion. Therefore, it is difficult for those RGB-based models to accurately distinguish the salient objects in complex scenarios, e.g., poor illumination conditions or clutter backgrounds (see Fig. 1). Being complementary to RGB images, thermal images, which capture the radiated heat of objects, can present silhouettes of salient objects clearly even in the case of insufficient illumination. However, they usually lack detailed visual cues as shown in the last column in Fig. 1. Therefore, it is desirable to combine the RGB images and thermal images to solve the salient object detection problem in complex visual scenes.

Essentially, with the rapid development of sensor technologies, the images in different modalities can be acquired easily.

Qiang Zhang is with the Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an Shaanxi 710071, China, and also with the Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China. Email: qzhang@xidian.edu.cn.

Tonglin Xiao, Nianchang Huang, and Dingwen Zhang are with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China. Email: tlxiao@stu.xidian.edu.cn, nchuang@stu.xidian.edu.cn, zdw@xidian.edu.cn.

Jungong Han is with Computer Science Department, Aberystwyth University, SY23 3FL, UK. Email: jungonghan77@gmail.com

Dingwen Zhang and Jungong Han are the corresponding authors.

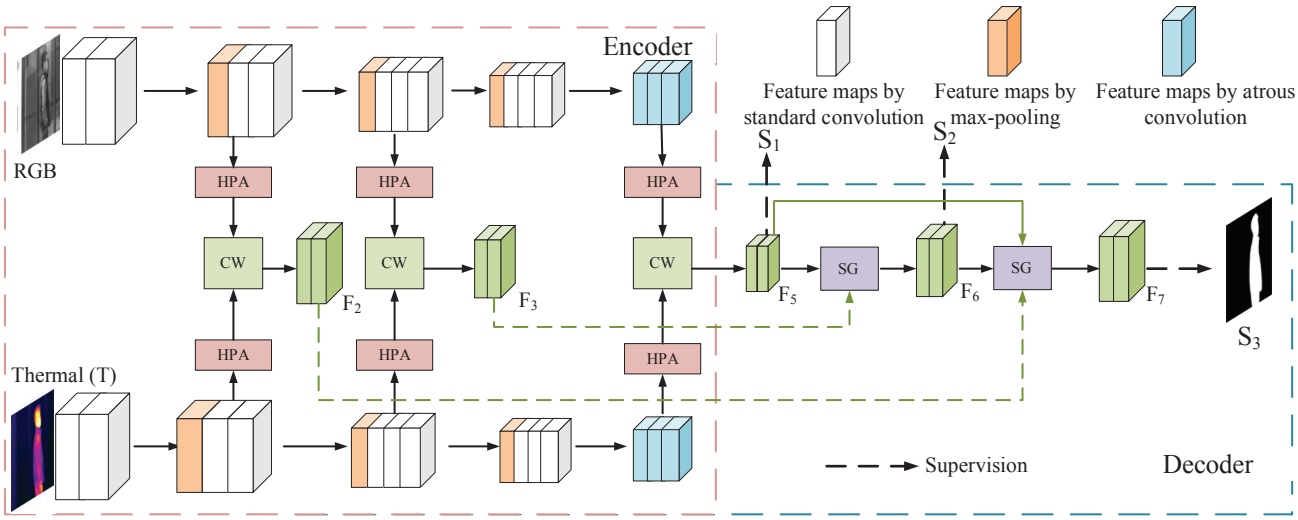


Fig. 2. The overall framework of our proposed model. Two feature extraction branches take an RGB image and a thermal image as input respectively and both of them are built on VGG-16 net. An HPA module is used to capture rich multi-scale contextual information for single modality images at different levels. Then the multi-scale multi-modality features are integrated by a CW module, which adaptively fuses multi-modality features by weighing their importance. Finally, the combined low-level RGB-T features are progressively integrated with the combined RGB-T global semantic features to predict fine saliency maps through a SG module, which employs global semantic features to supervise the passing of low-level details.

Furthermore, these multi-modality images have been extensively applied to different computer vision tasks. For instance, many RGB-D salient object detection works [24]–[34] have been proposed to work on depth images that contain affluent spatial structure and 3D layout information. However, depth data, in practice, are less robust to insufficient illumination and occlusion. A recent work [35] utilized thermal data to provide additional saliency information, where fusing multi-modal information at the saliency map level may not fully utilize the complementary information across RGB and thermal images. The latest RGB-T and RGB-D saliency detection methods [30], [32], [36] adopt feature fusion to combine multi-modality information. However, these methods fuse multi-modality features via simple concatenation or element-wise summation operation without considering the importance of the feature maps from different modalities. Such operations allows redundant or non-salient features to be involved, thus making the fusion process less complementary. Finally, most existing multi-modality salient object detection models may not take the multi-scale deep features into consideration, although multi-scale deep features have already been proved to be effective for the conventional salient object detection task [10], [37].

Considering those aforementioned issues, we propose a novel RGB-T saliency detection approach by revisiting feature fusion in the built Deep Convolutional Neural Network (DCNN) model. As shown in Fig. 2, the proposed DCNN model consists of a single-modality feature learning phase and a multi-modality feature learning phase. In this DCNN model, we study three feature fusion mechanisms for RGB-T saliency detection, including the multi-scale feature fusion, multi-modality feature fusion and multi-level feature fusion, respectively.

The multi-scale feature fusion is explored in the single-modality feature learning phase, where we propose a Hybrid Pooling-Atrous (HPA) module to capture multi-scale contextual information at each single-modality feature learning branch. This module introduces a cascade of hybrid atrous convolutional layers [38], where the gap between adjacent dilation rates is smaller, to alleviate the 'gridding' problem existing in the traditional atrous convolutional framework [39]. Thus, it can better aggregate contextual information in neighbor locations. Notice that, before each atrous convolutional layer in HPA, a max-pooling layer is employed to further gather the local information and enlarge the corresponding local receptive fields. Compared to the existing multi-scale feature extraction modules in [10], [37], our hybrid pooling-atrous module (HPA) is able to encode more representative multi-scale contextual information with stronger local information and better spatial consistency.

In multi-modality feature fusion, we design a Complementary Weighting (CW) module to effectively fuse complementary information from multi-modality features. Inspired by the existing works that use attention mechanisms [10], [40] to weigh and select features when fusing features at different levels, the proposed CW module applies the attention mechanism to adaptively fuse $multi\text{-}modality$ features. The attention mechanism is implemented by a series of standard convolutional layers with learnable parameters, which makes our entire system end-to-end trainable. Compared to the aforementioned multi-modality feature fusion strategy, the proposed CW module can adaptively fuse those important multi-modality information by learning the content-dependent weight vectors.

Having obtained the fused multi-modality features in different levels, the next step is to progressively combine the low-level RGB-T features with the high-level RGB-T semantic features to generate saliency maps with accurate semantics and fine boundaries. To this end, we design a Semantic Guidance (SG) module to screen the superfluous information in low-level

features, which is implemented by using the global semantic features from the deepest network layer to gate the forward flow of the low-level features so that the useful information is transmitted and superfluous information is abandoned. Then, the proposed RGB-T salient object detection model can achieve accurate saliency detection from those multi-modality complementary information.

In summary, this work has the following three-fold main contributions:

- By revisiting three informative feature fusion strategies in DCNN model, we propose a novel deep learning framework for RGB-T salient object detection.
- Novel network modules, i.e, the HPA module, CW module and SG module are proposed to learn rich contextual, complementary and semantic-aware features, respectively.
- Comprehensive experiments on the RGB-T saliency detection benchmark are conducted to demonstrate the superior capacity of our approach.

The rest of the paper is organized as follows. Section II briefly reviews some related works and Section III illustrates the proposed multi-modality salient object detection model in detail. Experimental results and conclusions are given in Section IV and Section V, respectively.

II. Related Work

A. RGB-based Salient Object Detection

Early RGB-based salient object detection methods are mainly designed on low-level hand-crafted features, which generally adopted the heuristic saliency priors for saliency detection, such as color contrast [41], [42], boundary priors [18], [43] and center priors [19]. Although these empirical saliency priors can improve saliency results for many images, they may fail when a salient object is similar to the background region in color distribution, off-center or significantly overlapped with the image boundaries.

Recently, many deep learning models with various convolutional neural network structures have been presented for saliency detection, which generally achieve better performance than the traditional hand-crafted features based methods. Wang *et al.* [9] proposed to utilize two branches of convolutional networks to combine local superpixel estimation and global proposal search for salient object detection. Li *et al.* [8] calculated the saliency value of each superpixel by learning its CNN contextual features. Although these models have achieved a good performance, the fully connected layers in these models decrease the computational efficiency and discard the spatial details. To address this issue, FCN based saliency detection models have become the mainstream because they may provide a pixel-wise prediction. Li *et al.* [44] proposed a multi-task deep saliency model based on the FCN with a global input and a global output. Some other salient object detection models [45]–[48] demonstrate that end-to-end deep networks are more effective in capturing local and global contextual information than image region-based models.

In addition, multi-scale features are another important factor in salient object detection tasks, because a diverse range of contextual information is conducive to enhance spatial consistency and accurately segment the salient region. The Pyramid Pooling Module (PPM) [49] and Atrous Spatial Pyramid Pooling (ASPP) [50] in the task of semantic segmentation adopted parallel pooling layers and atrous convolutional layers to obtain corresponding multi-scale features with various receptive field sizes, which have achieved great performance. In salient object detection methods, Lu *et al.* [10] employed several parallel atrous convolutional layers to extract multi-scale information at different levels. Hou *et al.* [11] utilized skip connections to combine high-level side-output and shallower side-output so as to obtain multi-scale feature maps at each level. Wang *et al.* [37] employed a pyramid pooling module for gathering multi-scale global contextual information.

Furthermore, there is an intrinsic problem in deep learning saliency detectors that low-level details will be lost with multiple levels of convolutional and pooling layers, leading to a coarse saliency map with blurry boundary. To tackle this problem, Zhang *et al.* [12] progressively combined high-level semantic information with low-level detail information to refine the sparse and irregular prediction maps. Alternatively, some salient object detectors refine the boundary of saliency maps by applying image regions [8], [9], [46] or superpixels [8]. For instance, Wang *et al.* [13] applied the superpixels generated by SLIC segmentation algorithm to refine boundary, and these superpixels can better keep boundary information of the salient object.

B. Multi-modality Salient Object Detection

In recent years, the multi-modal images have been applied to the salient object detection task. For example, many RGB-D saliency detection models have been presented, which can be categorized as three folds according to their fusion way: 1) input fusion; 2) feature fusion; and 3) result fusion. The input fusion models [24], [25] simply concatenate the RGB and depth image as a four-channel input without specific processing. The result fusion models computed the saliency maps for the RGB images and depth images separately and then adopted different fusion strategies to combine the two saliency maps, such as summation fusion [26], [27], multiplication fusion [28] or other fusion rules [29]. However, these input fusion or result fusion based RGB-D saliency detection models lack the interaction of multi-modality features at different levels, so they cannot effectively fuse the complementary information from RGB and depth images.

There are also many deep feature fusion models for RGB-D salient object detection. Chen *et al.* [30] leveraged two CNN models to extract information from RGB and depth images independently. And then multi-modality features at different levels were combined progressively by the complementary-aware fusion module. In addition, Zhu *et al.* [31] employed the encoder-decoder architecture as a master network for processing RGB information, and then the depth-based features acquired by a small sub-network were incorporated into the master network to enhance the robustness of the master network. Although these RGB-D salient object detection models

[31], [32] have achieved the state-of-the-art performance, they might not work well for RGB-T salient object detection task, as evidenced by the experimental results in Section IV. The major reason lies in the natural difference between thermal images and depth images.

Compared with the above-mentioned saliency detection methods on RGB-D images, the saliency detection methods on RGB-T image pairs have not received much attention. Currently, the paired RGB-T images are mainly applied in pedestrian detection [51]–[53] to deal with some complex scenarios such as lack of illumination or occlusion. With the introduction of the KAIST Multispectral Pedestrian Detection dataset [54], there is a growing interest in multispectral pedestrian detection leveraging aligned RGB and thermal images. For instance, Liu *et al.* [53] designed different multi-branch fusion network architectures based on Faster R-CNN to fuse RGB and thermal information, including low-level feature fusion, middle-level feature fusion, high-level feature fusion and predicted result fusion, and the middle-level feature fusion achieved the best performance.

Some works have applied thermal images to salient object detection task. Li *et al.* [55] adopted Weighted Low-rank Decomposition (WELD) for grayscale-thermal foreground detection, which adaptively pursued the cross-modality low-rank representation. Li *et al.* [56] associated a weight with each modality to describe the reliability and integrated them into a graph-based manifold ranking algorithm to achieve adaptive fusion of multi-modality information for RGB-T saliency detection. These existing traditional methods are designed on low-level hand-crafted features. When compared with multi-level features learned from CNN, these hand-crafted features are less discriminative. Then Tu *et al.* [57] posed saliency detection to a graph learning problem and utilized hierarchical deep features to jointly learn saliency. Ma *et al.* [35] adaptively incorporated RGB and thermal saliency maps inferred from deep convolutional neural networks. However, these latest RGB-T approaches never make full use of rich contextual information in DCNN, so they still cannot achieve the state-of-the-art performance. Recently, Zhang *et al.* [36] proposed the first end-to-end RGB-T salient object detection framework. This method combined multi-modality image information through multi-level feature fusion, which can better take advantage of the multi-modality hierarchy deep features. However, this method still did not consider the reliability of feature maps from each modality as well as the multi-scale deep features. Therefore, existing RGB-T salient object detection methods still fail to make full use of rich contextual information in DCNN and effectively integrate multi-modality complementary information.

III. Proposed RGB-T Salient Object Detection Method

A. Architecture Overview

As shown in Fig. 2, the proposed RGB-T salient object detection framework is based on an encoder-decoder architecture, where the two backbone streams (i.e., RGB and thermal streams) act as an encoder for encoding RGB images and thermal images information and the decoder predicts saliency maps based on the fused multi-modality information. Meanwhile, the proposed framework mainly consists of three feature fusion components: multi-scale contextual information extraction, multi-modality feature combination and multi-level feature propagation.

The encoder of the proposed network adopts the VGG-16 net [58] as the backbone architecture on both RGB and thermal streams to extract informative features. We first make some modifications to the VGG-16 net. Specifically, all the fully-connected layers in the VGG-16 net are discarded as our task focuses on pixel-wise prediction. Secondly, in order to enlarge the spatial resolution of the predicted saliency map and prevent the loss of spatial details, we remove the last two max-pooling layers and replace the subsequent three convolutional layers by the atrous convolutional layers with a dilation rate of 2. In this way, the output saliency map is down-sampled by a factor of 8 compared to the size of the original images.

After the RGB-T image pairs are fed into two modified VGG-16 nets, three levels of single-modality deep features (i.e., 2-nd, 3-rd and 5-th levels, respectively) are obtained. Here, due to the strides in the last two max-pooling layers have been set to 1, we do not employ feature maps from the fourth convolutional block to take part in the reasoning process. We also discard feature maps from the first convolutional block because of the experimental results in Section IV, which show that it would lead to performance degradation when these feature maps are employed to refine the final prediction maps. Then, a proposed HPA module is applied to each level of the obtained deep features to capture multi-scale contextual information for each modality. Therefore, we obtain three levels of multi-scale features $\left\{\mathbf{F}_i^{RGB}|i=2,3,5\right\}$ for RGB images and $\left\{\mathbf{F}_i^{T}|i=2,3,5\right\}$ for thermal images, respectively, in this step. Secondly, the proposed CW module is introduced to adaptively fuse the multi-scale RGB and thermal feature maps, i,e., $\mathbf{F}_i^{RGB}$ and $\mathbf{F}_i^{T}$, at the same *i*-th level ($i=2,3,5$). Thus, we obtain the fused RGB-T feature maps $\{\mathbf{F}_i|i=2,3,5\}$. Thirdly, in the phase of decoding, the proposed SG module is employed to adaptively combine multi-level fused RGB-T feature maps for recovering boundary details. In this way, two sets of refined RGB-T feature maps $\{\mathbf{F}_i|i=6,7\}$ are further produced. Finally, the high-level and refined RGB-T feature maps $\{\mathbf{F}_i|i=5,6,7\}$ are used to obtain three corresponding prediction maps $\{\mathbf{S}_j|j=1,2,3\}$ at the refinement stage for stage-wise supervisions. Here , all of the prediction maps have been resized to the same spatial resolutions as the input images by using the bilinear interpolation. Meanwhile, $\mathbf{S}_3$ is taken as the final prediction. We will discuss the proposed approach in details in the following sections.

B. Multi-Scale Features Fusion with Hybrid Pooling-Atrous Module

As we know, deep features from multiple scales are indispensable for salient object detection as the interested object regions may have large variations in scales. To this end, the multi-scale contextual information is required to enhance the prediction accuracy. To achieve this goal, Wang *et al.*

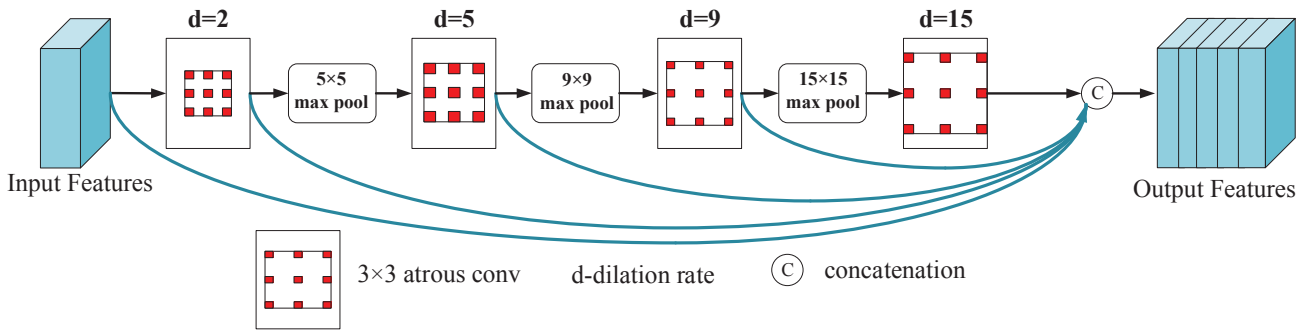


Fig. 3. The proposed HPA module used for the high-level semantic features in the fifth convolutional block. Each atrous convolutional layer is based on a 3×3 standard convolutional layer, and a max-pooling layer with a stride of 1 is added before each atrous convolutional layer except the first one.

[37] exploited the pyramid pooling before the final prediction layer to extract multi-scale features for saliency detection. However, the large scale of pooling may cause the loss of the detailed spatial information. In the semantic segmentation task, Chen *et al.* [50] proposed an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale contextual information by placing multiple atrous convolutional layers in parallel. Based on [50], Zhang *et al.* [10] exploited four parallel atrous convolutional layers with the same convolutional kernel size but different dilation rates to encode rich context information.

Usually, max-pooling layers are employed in DCNN to reduce the number of parameters and the amount of computation by gathering the local information into more salient cues. It can simultaneously obtain a larger receptive field and yield low-resolution feature maps. Using atrous convolutional layers to solve the problem of spatial resolution reduction is incurred by the repeated combination of max-pooling and down-sampling (striding) layers. It works by inserting zeros between kernel weights. Although the receptive field of the kernel increases, the number of filter parameters stays constant, which allows us to easily control the spatial resolution of feature maps.

However, there is a 'gridding' issue [39] in the atrous convolution-based frameworks. Specifically, to enlarge the receptive fields of a convolutional kernel, zeros are padded between two adjacent weights in the kernel. This means that only pixels corresponding to non-zero weights can be sampled during the calculation. So atrous convolutional operation may lose some useful neighboring information and destroy the informative spatial consistency. This issue gets worse with the increase of the dilation rate. The convolutional kernel is too sparse to cover any relative neighboring information for the sampled pixels. Although a larger receptive field is obtained, the sampled pixels cannot represent their relative local areas, which hurts the representative capacity of the deep features extracted at various scales.

To address this problem, we propose the HPA module to learn more representative multi-scale RGB and thermal contextual information. The HPA module contains a series of cascaded atrous convolutional layers for denser pixel sampling, and a max-pooling layer is added before each atrous convolutional layer, except the first one, to further gather local information around pixels with non-zero weights. In this way, the HPA module can enhance its representation capacity and enlarge the local receptive fields without introducing extra network parameters.

Specifically, the proposed HPA module consists of four atrous convolutional layers and three max-pooling layers. The kernel sizes of the three max-pooling layers are set as the same as the dilation rates of their next adjacent atrous convolutional layers. As a result, the next adjacent atrous convolutional layers can just cover a full local region even if zero weights are involved. For the high-level (i.e., the 5-th level) semantic features, this module consists of four 3×3 atrous convolutional layers with the dilation rates of 2, 5, 9, 15, respectively, to capture global semantic features. When learning multi-scale low-level (i.e, the 2-nd and 3-rd levels) features, the dilation rates of the four atrous convolutional layers are set to 1, 3, 5, 7, respectively, since there is no need to seek global semantic information anymore. By applying the proposed HPA module on deep features at these levels, multi-scale RGB features $\left\{\mathbf{F}_i^{RGB}|i=2,3,5\right\}$ and multi-scale thermal features $\left\{\mathbf{F}_i^{T}|i=2,3,5\right\}$ are acquired.

The concrete architecture of the proposed HPA module used for the high-level semantic features in the fifth convolutional block is illustrated in Fig. 3. The four atrous convolutional layers in HPA module produce various scales of feature maps. These feature maps are further integrated by skip connection and 1×1 convolutional operation. To avoid large-scale parameters, the size of output dimensionality in each atrous convolutional layer is set to 128, and the channel number of the final output feature maps of HPA module is also reduced to 128 by a 1×1 convolutional layer.

By cascading several atrous convolutional layers and max-pooling layers, the proposed HPA module can capture representative multi-scale contextual information with lager receptive fields and stronger local information. An example is seen in Fig. 4. Let $R_{K,d}$ denotes the receptive field of a layer with kernel size K and dilation rate d, and R_K denotes the receptive field of a max-pooling layer with kernel size K. R denotes the total receptive filed. As shown in Fig. 4 (a), in a two-dimension case, only 9 pixels are explored when using a layer with a dilation rate of 5. The corresponding receptive filed is:

$$R = R_{3,5}. \tag{1}$$

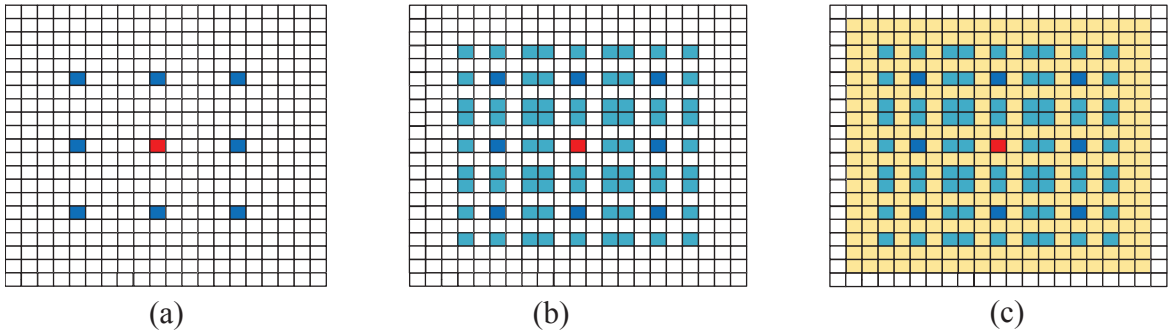


Fig. 4. Pixels sampling results with different atrous convolutional layer settings. (a) Only an atrous convolutional layer with a dilation rate of 5 is used. (b) Stacking an atrous convolutional layer with a dilation rate of 2 before the atrous layer with a dilation rate of 5. (c) Stacking an atrous convolutional layer with a dilation rate of 2 and a 5×5 max-pooling layer with a stride of 1 before the atrous convolutional layer with a dilation rate of 5. The blue and golden grids denote the pixels that contribute to the final calculation of the center pixel (marked in red).

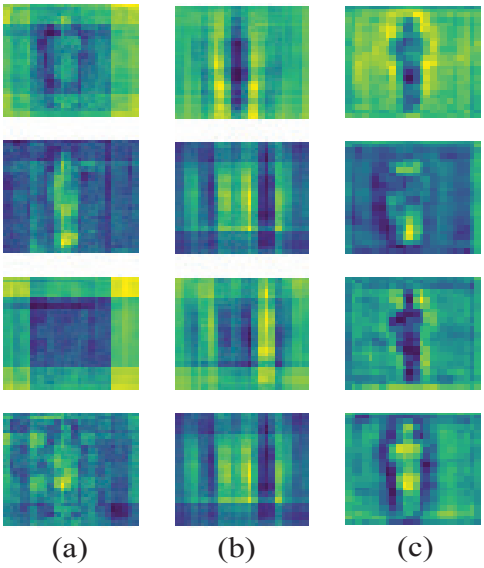


Fig. 5. Feature maps obtained by different atrous convolutional layer settings. (a) Only an atrous convolutional layer with a dilation rate of 5 is used. (b) Stacking an atrous convolutional layer with a dilation rate of 2 before the atrous layer with a dilation rate of 5. (c) Stacking an atrous convolutional layer with a dilation rate of 2 and a 5×5 max-pooling layer with a stride of 1 before the atrous convolutional layer with a dilation rate of 5.

However, when an atrous convolutional layer with a dilation rate of 2 is employed before the layer with the dilation rate of 5, 81 pixels will contribute to the final calculation as shown in Fig. 4 (b), and the corresponding receptive filed is [38]:

$$R = R_{3,5} + R_{3,2} - 1. \tag{2}$$

As a result, these cascaded atrous convolutional layers can sample pixels in a much denser way, thus increasing the correlation of pixels with non-zero weight, and reducing the loss of local information to some extent. If a max-pooling layer with a stride of 1 is further added between the two atrous convolutional layers, as illustrated in Fig. 4 (c), a larger receptive field would be obtained:

$$R = R_{3,5} + R_{3,2} + R_5 - 2. \tag{3}$$

These sampled pixels (dark blue grids) are more representative for they are selected from a full local area when a 5×5 max-pooling layer is added. With the dilation rate becoming lager, the receptive field of atrous convolutional kernel can still cover a full local region for each non-zero weights. Fig. 5 shows feature maps that are obtained by the HPA module under different settings. It can be observed that both the cascading operation for the atrous convolutional layers and the introduction of max-pooling layers contribute to the alleviation of the 'gridding' problem. Compared with the classic PPM [49] and ASPP [50], the proposed HPA module can also encode more representative multi-scale contextual information with stronger local information and better spatial consistency, which will be further verified in the experimental part.

C. Multi-Modality Feature Fusion with Complementary Weighting Module

To fuse multi-modality features from the RGB and thermal images, the most straightforward way is using element-wise summation operation. However, element-wise summation lacks cross-modal interactions and cannot leverage multi-modality complementary information efficiently. In recently published RGB-T and RGB-D salient detection methods [30], [32], [36], the multi-modality features fusion can also be implemented by the concatenation operation. However, this strategy still fails to consider the reliability of the features from different modalities. Furthermore, neither of the above two fusion strategies takes the content dependency of the multi-modality data into consideration. In this work, we propose a novel CW module to effectively fuse multi-modality features for integrating the RGB-T complementarities in a more effective manner.

For the i-th level, as shown in Fig. 6(c), the proposed CW module concatenates the integrated multi-scale feature maps from RGB branch $\left(\mathbf{F}_i^{RGB}\right)$ and thermal branch $\left(\mathbf{F}_i^{T}\right)$ as input maps. It consists of two convolutional layers and a softmax layer. The first convolutional layer has 128 filters with kernel sizes 3×3 and the second convolutional layer has 2 filters with the same kernel sizes to get two feature maps. The two convolutional layers provide a chance for cross-modal feature interaction, and then we obtain two-channel weight maps $\mathbf{Z}_i$:

$$\mathbf{Z}_i = Conv\big(Cat\big(\mathbf{F}_i^{RGB}, \mathbf{F}_i^{T}\big); \theta\big), \tag{4}$$

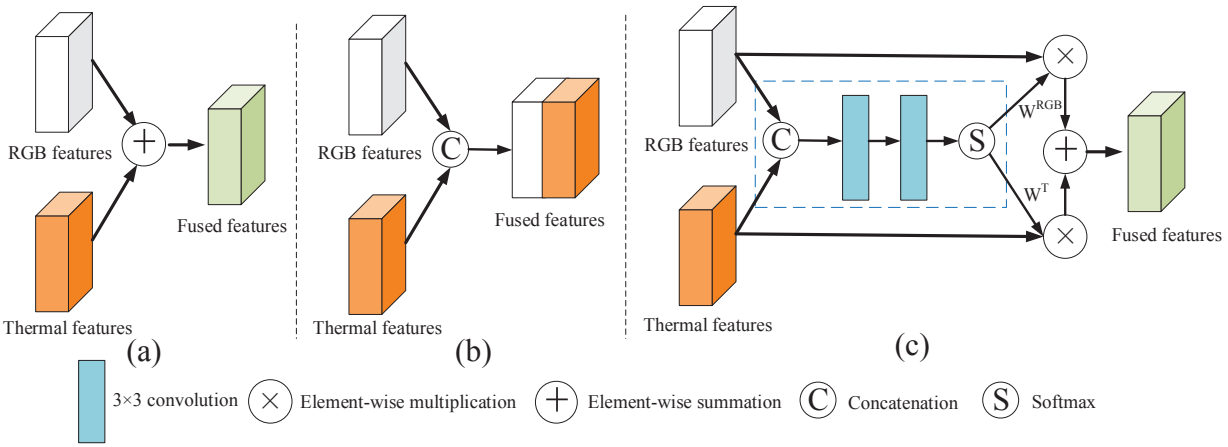


Fig. 6. Existing multi-modality feature fusion methods and the proposed multi-modality feature fusion module. (a) Multi-modality feature fusion by element-wise summation operation. (b) Multi-modality feature fusion by concatenation operation. (c) The proposed CW module. The CW module (blue dotted box) takes the multi-scale feature maps from HPA module as input and produces a multi-modality image content-dependent weight map that has the same spatial size as the input feature maps for each modality. The weight maps are able to weigh the importance of feature maps from RGB and thermal encoding branches and then adaptively fuse them.

where $Conv(*;\theta)$ denotes two convolutional layers, and θ denotes the parameters of the two convolutional layers. Then, a softmax layer is used to regularize the values of different locations in the feature maps to $[0,1]$. Finally, the two-channel weight maps are split into two weight maps, i.e., one weight map (denoted as $\mathbf{W}_i^{RGB}$) from the first channel of $\mathbf{Z}_i$ for selecting the features extracted from RGB images and another weight map (denoted as $\mathbf{W}_i^T$) from the second channel of $\mathbf{Z}_i$ for selecting the features extracted from thermal images. Finally, by virtue of the generated weight maps $\mathbf{W}_i^{RGB}$ and $\mathbf{W}_i^T$, the fused RGB-T feature maps $(\mathbf{F}_i)$ are calculated by:

$$\mathbf{F}_i = \widehat{\mathbf{W}}_i^{RGB} \otimes \mathbf{F}_i^{RGB} + \widehat{\mathbf{W}}_i^{T} \otimes \mathbf{F}_i^{T}, \tag{5}$$

where $\otimes$ is element-wise multiplication. The attention weight for the RGB features at the location (x,y) is computed by:

$$\widehat{\mathbf{W}}_i^{RGB}(x,y) = \frac{e^{\mathbf{W}_i^{RGB}(x,y)}}{e^{\mathbf{W}_i^{RGB}(x,y)} + e^{\mathbf{W}_i^{T}(x,y)}}, \tag{6}$$

where $\mathbf{W}_i^{RGB}(x,y)$ and $\mathbf{W}_i^{T}(x,y)$ are the corresponding weights for RGB feature maps and thermal feature maps, respectively. The attention weight for thermal feature at the location (x,y) is then obtained by:

$$\widehat{\mathbf{W}}_i^{T}(x,y) = 1 - \mathbf{W}_i^{RGB}(x,y). \tag{7}$$

By applying the proposed CW module on these integrated multi-scale RGB and thermal feature maps at the 2-nd, 3-rd and 5-th levels, respectively, multi-level fused RGB-T feature maps $\{\mathbf{F}_i | i = 2,3,5\}$ are thus acquired.

We further visualize some weight maps produced by the proposed CW module in Fig. 7, which clearly shows that our proposed CW module is able to learn some reliable weight maps for fusing the complementary information from different modalities. In this way, the weight maps can decide how much attention to pay to features from different modalities and different locations in a global view. Thus, the CW module can produce multi-modality image content-dependent weight maps for adaptively fusing multi-modality features.

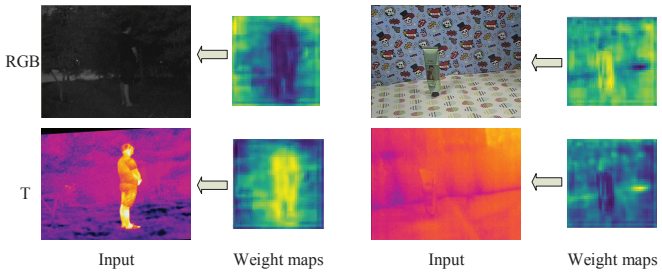


Fig. 7. Weights maps produced by CW module for multi-modality feature maps in the fifth level. The first image pairs show a situation of insufficient illumination, and the weight map for the thermal image has stronger values in the saliency region. For the second image pairs, when the thermal image cannot provide discriminative visual information, the weight map for RGB image shows higher reliability in the foreground location.

It should also be noted that the proposed CW module is a weighted averaging based one. The traditional element-wise summation and concatenation fusion strategies can be seen as two special cases of our proposed CW module. Specifically, element-wise summation can be seen as a weighted-averaging fusion strategy with weights that are all 1s. Concatenation followed by convolution operations can be seen as a weighted-averaging fusion strategy with weights that have been pre-trained and keep fixed during the subsequent fusion. While the weights in our proposed fusion strategy are dependent on the image contents.

D. Multi-Level Feature Fusion with Semantic Guidance Module

In the DCNN based models, the features extracted from deeper layers typically carry more global contextual information and are more likely to locate the salient object accurately, while the features extracted from shallower layers contain more spatial details, but may be less discrimination. So it's necessary to achieve saliency detection with hierarchical features. Some existing salient object detection methods [11],

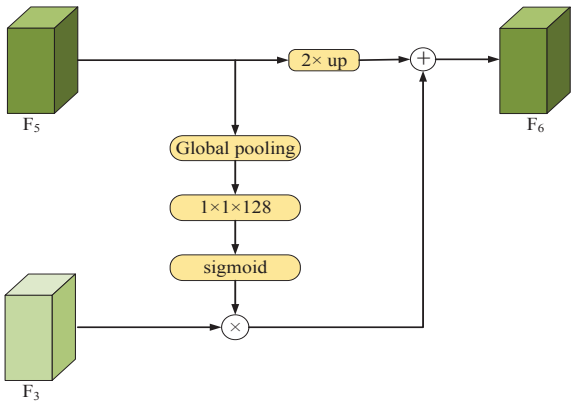


Fig. 8. The structure of the proposed SG module in the first stage of the refinement network. The fused high-level RGB-T semantic feature maps $\mathbf{F}_5$ act as a guidance for the pass of the fused low-level RGB-T feature map $\mathbf{F}_3$. After that, $\mathbf{F}_5$ is reused to filter out ambiguous information in $\mathbf{F}_2$. $\{\mathbf{F}_i|i=2,3,5\}$ are fused RGB-T feature maps acquired by fusing output feature maps from the HPA module in RGB branch and thermal branch.

[12], [46] proposed to combine multi-level convolutional features to obtain saliency maps with finer boundaries. However, these methods directly fuse multi-level features by cross-channel concatenation operation, which ignores the issue that the shallower features may contain superfluous information. This may lead to performance degradation or totally wrong prediction.

To filter out the misleading information when integrating multi-level features, we propose a progressive refinement network with a SG module to recover boundary details, which employs the final global semantic feature maps as a guidance to the fusion of the high-level and low-level RGB-T features. As we know, the global semantic feature maps obtained from the last network layer are more discriminative. Thus, it could be used to weigh the channels of shallower features and select more useful spatial details or screen the interference information. In this way, we acquire semantic-aware low-level features.

To implement the aforementioned network module, we progressively combine the fused low-level RGB-T feature maps $\{\mathbf{F}_i|i=2,3\}$ with the fused high-level RGB-T feature maps $\mathbf{F}_5$ to recover the spatial details. The refinement process consists of two stages. In the first stage, the fused low-level RGB-T feature maps $\mathbf{F}_3$ are selectively added to the high-level semantic RGB-T features $\mathbf{F}_5$ with the guidance of $\mathbf{F}_5$. A set of refined RGB-T feature maps $\mathbf{F}_6$ are obtained. In the second stage, the fused low-level RGB-T feature maps $\mathbf{F}_2$ are adaptively combined with the refined RGB-T feature maps $\mathbf{F}_6$ with the guidance of $\mathbf{F}_5$, and then another set of refined RGB-T feature maps $\mathbf{F}_7$ are obtained. This module is inspired by the "Squeeze-and-Excitation" network (SENet) [59], which adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. Different from the SENet, our module employs global semantic feature maps to recalibrate the local information passed on.

Fig. 8 illustrates the detailed structure of the SG module in the first stage. Specifically, $\mathbf{F}_5$ is first passed through a squeeze operation, which produces a channel descriptor by aggregating feature maps across their spatial dimensions. This is achieved by using global average pooling. Then a 1×1 convolutional layer with $C=128$ filters is employed to learn a nonlinear interaction among channels. The output vector $\mathbf{h}$ is formulated as

$$\mathbf{h}=\mathbf{W}*GAP\big(\mathbf{F}_5\big)+\mathbf{b}, \tag{8}$$

where $*$ denotes convolutional operation. $GAP\,(\cdot)$ denotes the global average pooling function. $\mathbf{W}\in R^{C\times1\times1\times C}$ represents C convolutional filters, where each of the filters with dimension of $1\times1\times C$, and $\mathbf{b}\in R^{C}$ is the bias parameter. Finally, a sigmoid layer is used to regularize the corresponding channel weights to the range of $[0,1]$. The corresponding weight w^k for the k-th channel is defined as

$$\mathrm{w}^k=\frac{1}{1+e^{-\mathrm{h}^k}}, \tag{9}$$

where h^k is the k-th element of $\mathbf{h}$ and $\mathbf{w}\in R^{1\times1\times C}$ is the final guidance weight vector for the channels in low-level feature maps. In summary, the whole process in Fig. 8 can be formulated as:

$$\mathbf{F}_6^k=\mathbf{F}_3^k\times\mathrm{w}^k+UP\big(\mathbf{F}_5^k\big)_2, \tag{10}$$

where k denotes the k-th feature channel, and $UP\,(\cdot)_2$ denotes upsampling feature maps by a factor of 2. $\times$ denotes the channel-wise production.

The weights in $\mathbf{w}$ are reused to combine $\mathbf{F}_6$, $\mathbf{F}_2$ and $\mathbf{F}_5$ in the second stage. The high-level semantic feature maps $\mathbf{F}_5$ are employed again through a shortcut connection to mitigate the potential interference brought by low-level feature maps in this stage. After all, the semantic features dominate the refinement network. Accordingly, the refined feature maps $\mathbf{F}_7$ are obtained:

$$\mathbf{F}_7^k=\mathbf{F}_2^k\times\mathrm{w}^k+UP\big(\mathbf{F}_6^k\big)_2+UP\big(\mathbf{F}_5^k\big)_4. \tag{11}$$

With the semantic guidance, the superfluous information in low-level layers will be filtered out and the refinement network can produce more accurate saliency prediction with semantic-aware spatial details. It should also be noted that the three feature maps $\{\mathbf{F}_i|i=5,6,7\}$ have the same channel dimensionality of 128.

E. Stage-wise Intermediate Supervision

To promote the training of the proposed network, deep supervisions [60] are adopted in multiple refinement stages $\{\mathbf{F}_i|i=5,6,7\}$. Specifically, a two-channel 3×3 convolutional layer is applied upon the feature maps $\{\mathbf{F}_i|i=5,6,7\}$ at each refinement stage to obtain the corresponding prediction map. As a result, three saliency maps (denoted as $\{\mathbf{S}_j|j=1,2,3\}$, respectively) are obtained, where $\mathbf{S}_3$ is seen as the final prediction. Here, all of the prediction maps have been resized to the same spatial resolutions as the input images by using the bilinear interpolation.

Let $\mathbf{Y}(x,y)\in[0,1]$ denote the ground-truth mask, where (x,y) is the pixels location. The cross-entropy loss

$\{l_j|j=1,2,3\}$ between the prediction map $\mathbf{S}_j$ and the ground-truth $\mathbf{Y}$ is defined as:

$$\begin{aligned} l_j = &- \sum_{(x,y)} \mathbf{Y}(x,y) \log \mathbf{S}_j(x,y) \\ &+ (1-\mathbf{Y}(x,y)) \log (1-\mathbf{S}_j(x,y)). \end{aligned} \tag{12}$$

Thus, the final loss function of the whole RGB-T salient object detection network is:

$$L_{final} = \gamma_1 l_1 + \gamma_2 l_2 + \gamma_3 l_3, \tag{13}$$

where γ_1, γ_2 and γ_3 denote the hyper-parameters for the three stage-wise losses, which are all set to 1 as used in [12] for simplicity and fair comparisons. With the SG module and stage-wise intermediate supervisions, the low-level details and high-level semantic information can be effectively fused, and the boundary details of the salient object may also be recovered accurately. As a result, we can produce a refined prediction map with precise object boundaries and fine spatial consistency.

IV. Experiments

A. Datasets

We conduct our experiments on a RGB-T saliency detection benchmark dataset [56], which contains 821 aligned RGB-T image pairs. To enlarge the size of the labeled RGB-T dataset, we follow [55] to select 539 aligned image pairs with ground truth annotations from a grayscale-thermal dataset that includes 25 small sets for moving foreground detection. The images in both of the two datasets are with complex backgrounds or lack of illumination. Similar to that in [55], the whole RGB-T dataset is divided into two parts by randomly sampling 621 and 385 RGB-T image pairs from the RGB-T benchmark and grayscale-thermal datasets as the training set and the remaining image pairs are for testing. We also employed 2000 images from MSRA-B [63] when pre-training the RGB feature encoding branch. To make the model robust, we augment the training set by horizontal flipping, rotating, random brightness and contrast changing.

B. Evaluation Metrics

We adopt the standard metrics (Precision-Recall (PR) curve, F-measure score (F_β) and Mean Absolute Error (MAE)) to evaluate the proposed method. The Precision value is the ratio of ground truth salient pixels in the predicted salient region. And the Recall value is defined as the percentage of the detected salient pixels and all ground truth area. The PR curve is computed by binarizing the saliency maps under different probability thresholds ranging from 0 to 1 and comparing to the ground truth. The formulation of F-measure is

$$F_\beta = \frac{\left(1+\beta^2\right) \times \mathit{precision} \times \mathit{recall}}{\beta^2 \times \mathit{precision} + \mathit{recall}}, \tag{14}$$

where $\beta^2 = 0.3$ to emphasize precision more than recall as suggested in [43]. We report the mean F-measure computed from the PR curve. MAE [64] is computed as the average pixel-wise absolute difference between the estimated saliency map $\mathbf{S}$ and its corresponding ground truth $\mathbf{Y}$. It can be defined as:

$$\mathit{MAE} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\mathbf{S}(x,y) - \mathbf{Y}(x,y)|, \tag{15}$$

where W and H are the width and height of a given image.

C. Implementation Details

We utilize the popular TensorFlow framework [65] to implement the proposed network. An NVIDIA GTX 1080 Ti GPU is used for training and testing. During training, the weights of the first 13 convolutional layers are initialized by the VGG-16 net [58]. For other convolutional layers, including HPA module, CW module and SG module, we initialize the weights randomly with a truncated normal, and the biases are initialized to 0. The up-sampling operation is conducted by the bilinear interpolation in the proposed model. Our training process can be divided into two stages. In the first stage, we train two independent feature encoding branches for RGB and thermal images, respectively, each of which contains 13 convolutional layers and four atrous convolutional layers. The training process of each branch takes about 3 hours for 20 epochs with the batch size of 1. In the second stage, we fine-tune the whole encoder-decoder network on the aligned RGB-T image pairs, which takes about 6 hours to converge after 10 epochs. The Adam optimizer [66] is used in both stages with an initial learning rate of 10^{-5}. The test time for each RGB-T image pair is merely 0.056s.

D. Comparison with State-of-the-art methods

We compare the proposed multi-modality saliency detection model with 11 state-of-the-art methods on two datasets, including 5 deep learning based RGB saliency detection methods (BMPM [10], DSS [11], Amulet [12], UCF [61], and CPD [62]), 4 RGB-T saliency detection approaches (MRCMC [56], MFSR [35], CGL [57], and FMCF [1] [36]) and 2 latest RGB-D salient object detection models (PDNet [31] and TSAA [32]). To better verify the superiority of the proposed model, as shown in Table I, these state-of-the-art methods are compared under different settings of the input modality, i.e., only taking RGB images, only taking thermal images and taking RGB-T images as inputs, respectively. Specifically, those RGB saliency detection models are modified into thermal saliency detection models by replacing the RGB images with thermal images as inputs. The procedure of converting the RGB saliency detection model into an extended RGB-T model is described as follows. First, their proposed networks are taken as the backbones of the RGB and thermal branches, respectively. Then, the output features from the last convolutional layers of the RGB and thermal branches are concatenated. Finally, the concatenated features are fed into the saliency prediction layer to obtain the saliency map. Here, the saliency prediction layer contains standard convolutional layers and a Sigmoid activation layer. Those RGB-T saliency detection models are modified into RGB (thermal) saliency detection

[1]FMCF has been re-trained in our training sets for fair comparisons.

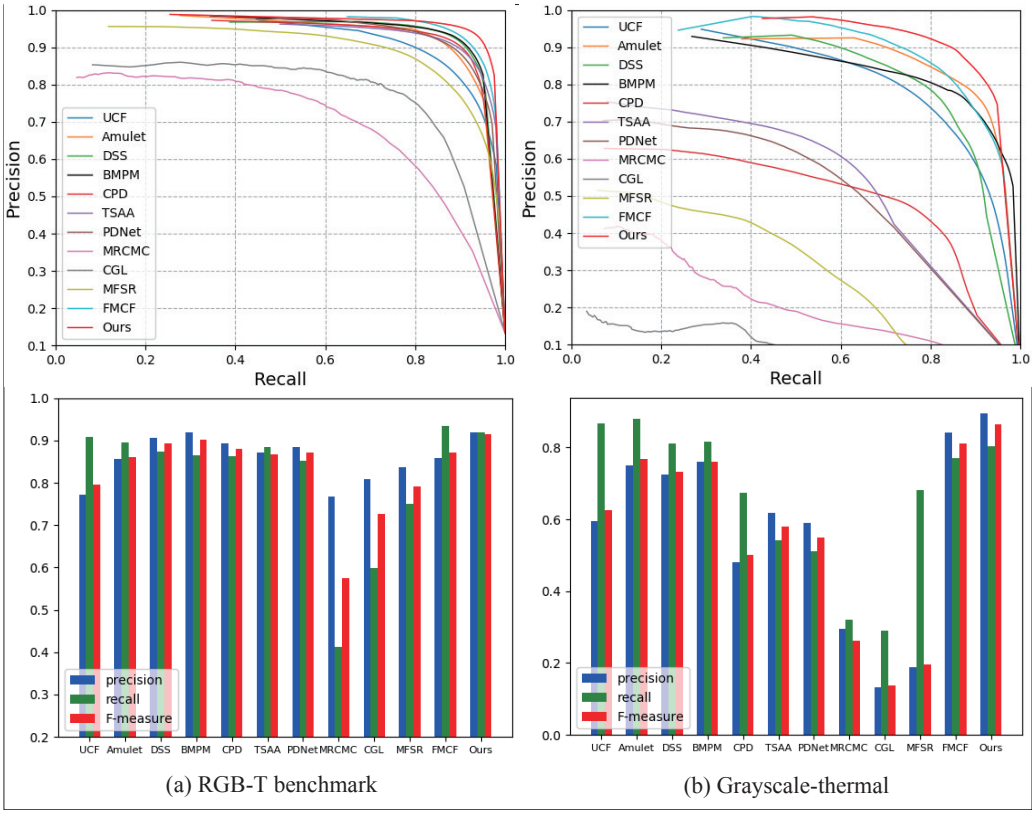


Fig. 9. The quantitative comparison of the proposed method and other state-of-the-art methods on the RGB-T benchmark dataset and the Grayscale-thermal dataset. The first row shows the PR curves. The second row shows the average precision, recall, and F-measure scores of different methods, respectively.

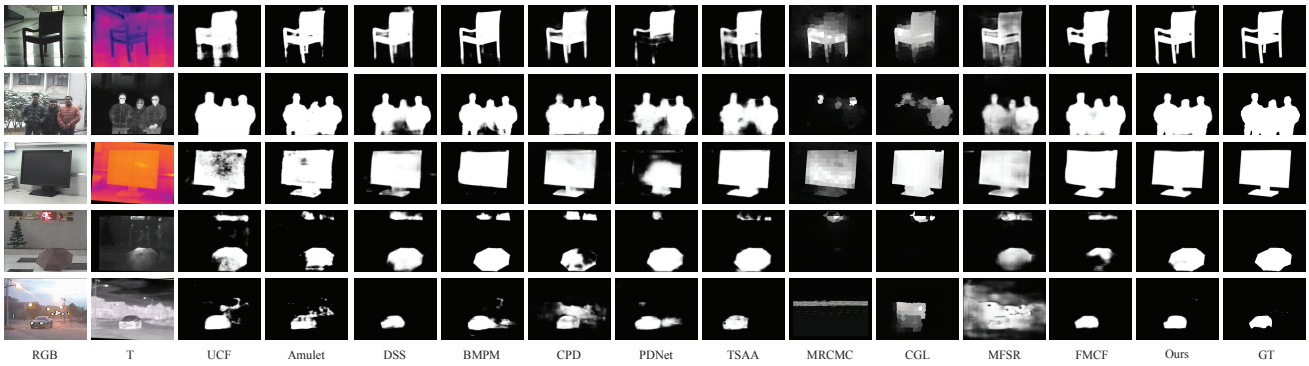


Fig. 10. Visual comparisons among the saliency detection results of the state-of-the-art methods in general scenarios.

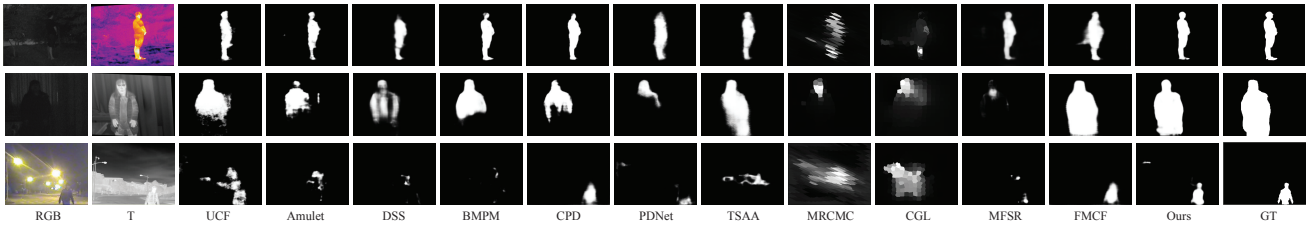


Fig. 11. Visual comparisons among the saliency detection results of the state-of-the-art methods in scenarios with insufficient illumination.

models by replacing the thermal (RGB) images with the RGB (thermal) images as inputs. Similarly, those RGB-D saliency detection models are modified into RGB (RGB-T) saliency detection models by replacing the depth images with RGB (thermal) images as inputs. Those RGB-D models are modified into thermal saliency detection models by replacing both the RGB images and depth images with thermal images. For fair comparisons, we train these state-of-the-art methods with the

TABLE I
QUANTITATIVE EVALUATION ON DIFFERENT SALIENCY DETECTION METHODS IN TERMS OF MEAN F-MEASURE (LARGER IS BETTER), MAE (SMALLER IS BETTER) AND RUNTIME (IN SECONDS). THE BEST THREE RESULTS ARE SHOWN IN RED, BLUE AND GREEN.

Methods	RGB-T benchmark						Grayscale-thermal						Runtime(s)
	RGB		T		RGB-T		RGB		T		RGB-T		
	F_β	*MAE*	F_β	*MAE*	F_β	*MAE*	F_β	*MAE*	F_β	*MAE*	F_β	*MAE*	
Amulet [12]	0.738	0.068	0.797	0.049	0.870	0.033	0.467	0.066	0.635	0.040	0.769	0.015	0.161
UCF [61]	0.741	0.059	0.700	0.066	0.794	0.047	0.465	0.056	0.514	0.046	0.626	0.031	0.334
DSS [11]	0.785	0.043	0.833	0.042	0.896	0.026	0.603	0.026	0.581	0.070	0.737	0.023	0.051
BMPM [10]	0.858	0.036	0.819	0.043	0.905	0.028	0.699	0.029	0.576	0.045	0.759	0.023	0.581
CPD [62]	0.793	0.034	0.840	0.033	0.881	0.028	0.199	0.114	0.451	0.112	0.500	0.103	0.166
PDNet [31]	0.833	0.043	0.796	0.051	0.870	0.032	0.489	0.035	0.488	0.036	0.550	0.034	0.113
TSAA [32]	0.803	0.048	0.680	0.076	0.867	0.029	0.461	0.034	0.491	0.043	0.579	0.028	0.135
MRCMC [56]	0.537	0.109	0.554	0.124	0.574	0.115	0.132	0.095	0.221	0.105	0.263	0.096	1.891
CGL [57]	0.706	0.084	0.668	0.098	0.726	0.088	0.112	0.132	0.123	0.143	0.138	0.136	2.332
MFSR [35]	0.416	0.161	0.554	0.124	0.791	0.062	0.180	0.164	0.175	0.135	0.196	0.125	0.183
FMCF [36]	0.759	0.068	0.778	0.048	0.880	0.025	0.573	0.032	0.436	0.052	0.811	0.016	0.110
Ours	0.875	0.035	0.820	0.045	0.915	0.021	0.665	0.027	0.574	0.069	0.833	0.011	0.056

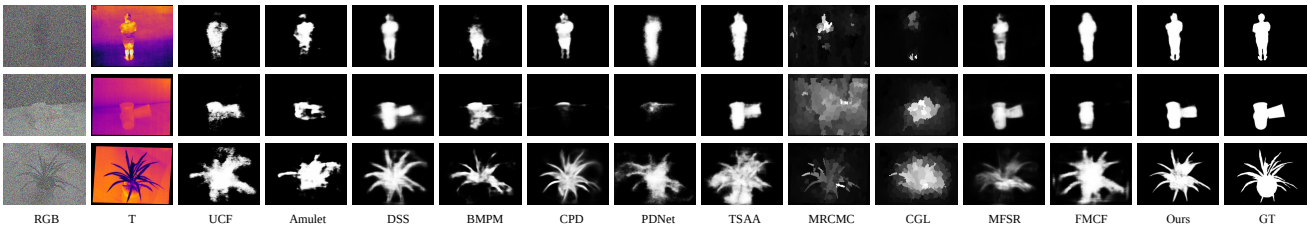


Fig. 12. Visual comparisons among the saliency detection results of the state-of-the-art methods in scenarios with occluded appearances.

same two-stage strategy as that in our approach.

The quantitative comparison results are shown in Table I and Fig. 9. From the results reported in Table I, we can observe that all the RGB-T and extended RGB-T methods outperform the corresponding RGB or thermal saliency detection methods. This demonstrates the effectiveness of incorporating RGB and thermal information. For those RGB-T saliency detection models, our method consistently outperforms other approaches on both datasets in terms of all the metrics, which demonstrates the effectiveness of the proposed model. From the PR curves and F-measure scores shown in Fig. 9, we can observe that our approach achieves better results than other state-of-the-art RGB-T, extended RGB-T and RGB-D saliency detection methods with a large margin on the two RGB-T datasets. This indicates that our method can more effectively make use of the complementary information within RGB-T images than other models. This also indicates that directly employing RGB-D salient object detection models for RGB-T salient object detection is not suitable due to the differences between depth images and thermal images. Furthermore, as indicated in Table I, our method has the second highest computation efficiency among these models mentioned here.

Fig. 10 provides the visual comparisons of our method with the above-mentioned models. These images are selected from the two testing datasets. It can be observed that our methods can accurately detect multi-scale salient objects with stronger spatial consistency. Furthermore, many interference information belonging to the non-salient regions could be filtered out. Fig. 11 and Fig. 12 show the visual comparisons to the state-of-the-art methods in scenarios with some insufficient illumination and occluded appearances. As can be seen from these visual comparison results, the proposed model can achieve superior performance in these complex scenarios due to the fact that the multi-modality information can be effectively fused with the help of CW module, and the detected salient objects present clear boundaries through our refinement network.

TABLE II
QUANTITATIVE COMPARISONS OF THE PROPOSED MODEL WITH AND WITHOUT AN HPA, HA OR SOME OTHER MULTI-SCALE CONTEXTUAL FEATURE EXTRACTION MODULES (PPM AND PAM).

Metrics	baseline	w/ PPM	w/ PAM	w/ HA	w/ HPA
F_β	0.880	0.885	0.889	0.902	**0.904**
MAE	0.036	0.033	0.032	0.028	**0.027**
Runtime(s)	**0.036**	0.051	0.053	0.053	0.053

E. Analysis of the Proposed Method

The proposed method mainly consists of three modules, including HPA module for multi-scale contextual features

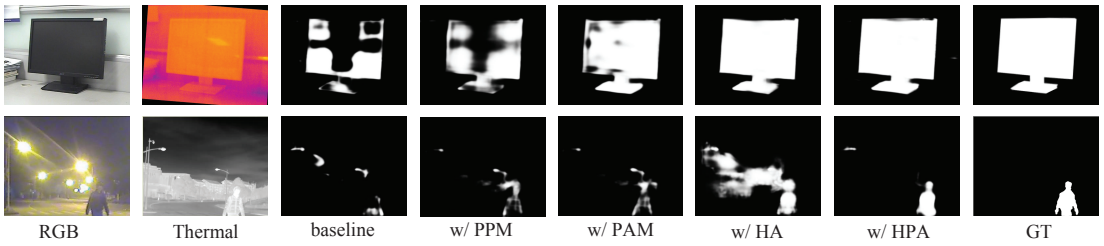


Fig. 13. Visual comparisons of the proposed model with and without an HPA, HA or some other multi-scale contextual feature extraction modules (PPM and PAM).

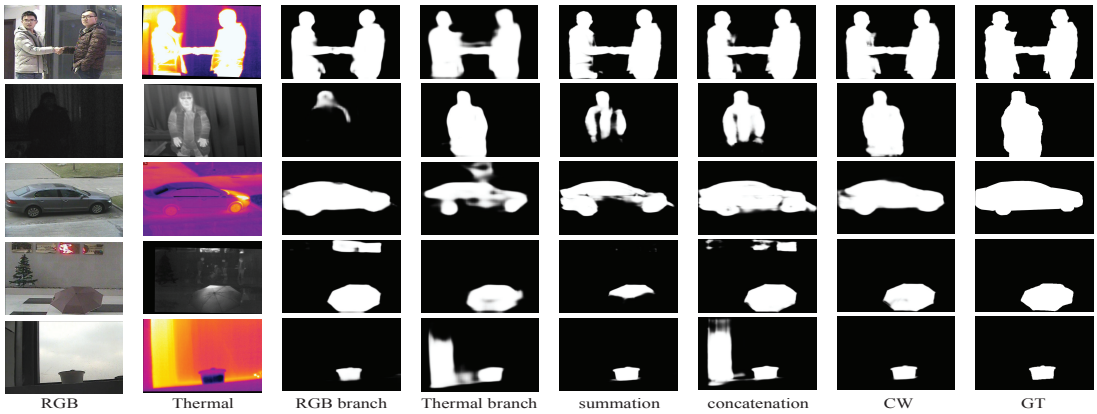


Fig. 14. Visual comparisons of the proposed model with different fusion approaches (CW, summation and concatenation).

TABLE III
QUANTITATIVE COMPARISONS OF DIFFERENT FUSION METHOD SETTINGS IN OUR MODEL.

Metrics	Summation	Concatenation	CW
F_β	0.902	0.904	**0.908**
MAE	0.030	0.027	**0.023**
Runtime(s)	**0.052**	0.053	0.055

extraction, CW module for multi-modality feature fusion and SG module for semantic-aware features. To explore the effectiveness of each module, we conduct a series of experiments on the RGB-T benchmark dataset [56]. First, to verify the validity of the HPA module, we built a baseline model by removing the HPA module and SG module from the proposed model and replacing CW module with concatenation operation. Then, we compare the proposed HPA module with some other most commonly used multi-scale contextual feature extraction modules (i.e., Pyramid Atrous Module (PAM) [10] and Pyramid Pooling Module (PPM) [37]) by incorporating those modules into the baseline model. In addition, we conduct a comparison between the HPA module and the HA module (i.e., removing the max-pooling layers from the HPA module). From Table II, we can observe that the proposed HPA module achieves significant improvements over PAM and PPM. In addition, the three modules, i.e., HPA, PAM and PPM, have almost the same computational complexity. This may owe to the fact that the pooling layer introduced in HPA module can strengthen the local information without involving additional parameters. Fig. 13 further illustrates that the proposed HPA module can enforce the spatial consistency within the salient objects and suppress the noise within the backgrounds, through capturing multi-scale contextual information with stronger local information. It also indicates that, compared with other multi-scale feature extraction modules, the proposed HPA module shows better robustness when the sizes of salient objects are various. This again demonstrates that introducing max-pooling layers between a series of cascaded atrous convolutional layers is effective to capture the multi-scale contextual information for salient object detections.

To highlight the CW module, we conduct comparisons with another two multi-modality feature fusion approaches: summation and concatenation. The visualization results are shown in Fig. 14 and the quantitative results, including mean F-measure, MAE and average computation time for each RGB-T image pair (runtime), are shown in Table III. From Table III and Fig. 14, we can observe that all of the three fusion methods can achieve good fusion results when both the input RGB and thermal images have good visual qualities. However, when one of the input images has poor visual quality,the summation fusion strategy would result in incomplete segmentation of the salient objects (e.g., the fifth column of Fig. 14). Under this situation, the concatenation fusion strategy may also introduce the non-salient information into the generated saliency maps (e.g., the sixth column of Fig. 14). In contrast, the proposed CW module can obtain better saliency detection results and outperform these two fusion strategies to obtain high quality saliency prediction. In addition, our fusion strategy introduces fewer extra computational costs compared with existing fusion

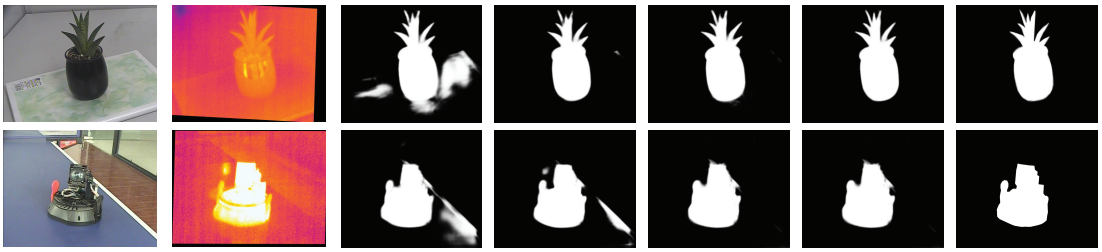


Fig. 15. Visual comparisons of the proposed model with and without SG module.

TABLE IV
QUANTITATIVE COMPARISONS OF THE PROPOSED MODEL WITH AND WITHOUT SG MODULE.

Metrics	w/o SG	SG(1,2,3)	SG(2,3,4)	SG(2,3)
F_β	0.908	0.909	0.914	**0.915**
MAE	0.023	0.023	0.021	**0.021**
Runtime(s)	**0.055**	0.063	0.061	0.056

TABLE V
SALIENCY DETECTION RESULTS OF THE PROPOSED MODEL UNDER DIFFERENT LOSS SETTINGS.

hyper-parameters (γ_1, γ_2, γ_3)	F_β	MAE	Training epoch of convergence
(1,1,1)	**0.915**	**0.021**	**10**
(1,0,0)	0.915	0.022	21
(1,0,1)	0.914	0.021	16
(1,0.5,0.5)	0.915	0.021	12

methods.

We embed SG module into the learning framework, and the comparison results are shown in Table IV and Fig. 15. The quantitative results verify the effectiveness of the proposed SG module. Meanwhile, it also indicates that, compared with w/o SG module, the proposed SG module only introduces a few more computational costs. Fig. 15 further illustrates the proposed SG module can effectively suppress noise introduced by features from the shallower levels.

As shown in Table IV, we further exploit which levels of the features should be employed in the proposed SG module for proper refinement setting. Here, SG(1,2,3) employs the feature maps in the first three convolutional blocks to refine the prediction maps and SG(2,3,4) employs feature maps in the 2-nd, 3-rd and 4-th convolutional blocks. Compared to SG(2,3), the introduction of the feature maps in the first block leads to performance degradation. This be due to that the feature maps in the first convolutional blocks carry too many spatial details. We can also see that using the feature maps in the 4-th convolutional block can not further improve the saliency detection performance. Therefore, we build the proposed module with HPA module, CW module and SG(2,3) module to achieve the new state-of-the-art results. Furthermore, replacing the VGG-16 net [58] with more recent CNN models as the backbone network may further boost the performance of the proposed model for RGB-T saliency detection. More specifically, the mean F-measure and MAE values are improved to (0.917, 0.021) by replacing the VGG-16 net with the VGG-19 [58] and it can be further improved to (0.918, 0.019) when replacing the VGG-16 net with the ResNet-50 [67].

Beside, we also perform some ablation analyses on the final loss by setting the three hyper-parameters (γ_1, γ_2, γ_3) in Eq. 13 to different values. The experimental results in Table V demonstrate that the performance of the proposed model is insensitive to the settings of these hyper-parameters. However, the settings of these hyper-parameters may affect the convergence speed of the proposed model. For example, when (γ_1, γ_2, γ_3) are set to (1,0,0), our network needs 21 epochs to converge. Whereas when (γ_1, γ_2, γ_3) are set to (1,1,1), our network only needs 10 epochs to converge. This can significantly reduce the total training time. Therefore, we set the three hyper-parameters (γ_1, γ_2, γ_3) to the same value 1 since the setting achieves the fastest convergence speed.

V. CONCLUSION

In this paper, we propose a novel end-to-end salient object detection method on RGB-T image pairs by revisiting feature fusion in the DCNN model. We first design an HPA module, which is composed of a cascade of atrous convolutional and max-pooling layers, to fuse more representative multi-scale contextual features with various receptive fields and stronger local information. Then, we propose a CW module to effectively fuse multi-modality complementary features at different levels, which allows us to diagnostically visualize the importance of features from different modalities. Finally, a multi-level feature fusion network branch is designed with SG module to produce saliency maps with fine boundaries. Comprehensive experiments on different datasets demonstrate the effectiveness of the proposed RGB-T salient object detection model.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 61773301 and 61876140, the China Postdoctoral Support Scheme for Innovative Talents under Grant No. BX20180236.

REFERENCES

[1] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 2–2.

[2] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2009.

[3] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 817–824.

[4] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 10, 2007.

[5] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 23–30.

[6] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3D object retrieval and recognition with hypergraph analysis," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4290–4303, 2012.

[7] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3586–3593.

[8] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[9] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.

[10] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1741–1750.

[11] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, pp. 815–828, 2019.

[12] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[13] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 121–134, 2017.

[14] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 8, pp. 1–14, 2019.

[15] R. Quan, J. Han, D. Zhang, F. Nie, X. Qian, and X. Li, "Unsupervised salient object detection via inferring from imperfect saliency models," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1101–1112, 2017.

[16] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2014.

[17] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[18] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2814–2821.

[19] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2043–2050.

[20] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 110–119.

[21] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended CSSD," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 717–729, 2015.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[23] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1395–1403.

[24] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4204–4216, 2017.

[25] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *European Conference on Computer Vision*, 2014, pp. 92–109.

[26] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *Proceedings of the IEEE International Conference on Digital Signal Processing*, 2014, pp. 454–458.

[27] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2625–2636, 2014.

[28] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of the International Conference on Internet Multimedia Computing and Service*, 2014, p. 23.

[29] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *Proceedings of the British Machine Vision Conference*, 2013, pp. 9–13.

[30] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[31] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2019, pp. 199–204.

[32] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[33] R. Huang, Y. Xing, and Z. Wang, "RGB-D salient object detection by a cnn with multiple layers fusion," *IEEE Signal Processing Letter*, vol. 26, no. 4, pp. 552–556, 2019.

[34] J. Zhao, Y. Cao, D. Fan, M. Cheng, X. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[35] Y. Ma, D. Sun, Q. Meng, Z. Ding, and C. Li, "Learning multiscale deep features and SVM regressors for adaptive RGB-T saliency detection," in *Proceedings of the International Symposium on Computational Intelligence and Design*, vol. 1, 2017, pp. 389–392.

[36] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level cnn features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2020.

[37] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4019–4028.

[38] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[39] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 1451–1460.

[40] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 714–722.

[41] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[42] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[43] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[44] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.

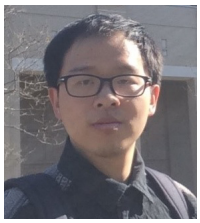
[45] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 660–668.

[46] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.
[47] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.
[48] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6609–6617.
[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.
[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[51] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *Proceedings of the European Symposium on Artificial Neural Networks*, 2016, pp. 509–514.
[52] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *Proceedings of the IEEE International Conference on Pattern Recognition*, 2016, pp. 621–626.
[53] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.
[54] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1037–1045.
[55] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 725–738, 2016.
[56] C. Li, G. Wang, Y. Ma, A. Zheng, B. Luo, and J. Tang, "RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Proceedings of the Chinese Conference on Image and Graphics Technologies*, 2018, pp. 359–369.
[57] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *arXiv preprint arXiv:1905.06741*, 2019.
[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[59] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
[60] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2015, pp. 562–570.
[61] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 212–221.
[62] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.
[63] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2010.
[64] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 733–740.
[65] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 770–778.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing, pattern recognition.

Tonglin Xiao received the B. E. degree from Xidian University, Xi'an, China, in 2017, and the M. S. degree from Xidian University, Xi'an, China, in 2020. He is currently studying at Xidian University. His current research interests include computer vision and machine learning.

Nianchang Huang received the B. S. degree and the M. S. degree from Qingdao University of Science and Technology, Qingdao, China, in 2015 and 2018. He is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. His research interests include deep learning and multimodal image processing in computer vision.



Dingwen Zhang received his Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently an associate professor in School of Machine-Electronical Engineering, Xidian University. From 2015 to 2017, he was a visiting scholar at the Robotic Institute, Carnegie Mellon University. His research interests include computer vision and multimedia processing, especially on saliency detection, video object segmentation, and weakly supervised learning.

Jungong Han is currently a Full Professor and Chair in Computer Science at Aberystwyth University, UK. His research interests span the fields of video analysis, computer vision and applied machine learning. He has published over 180 papers, including 40+ IEEE Trans and 40+ A* conference papers.